Collapse and Phase Transition in Deep Learning

Liu Ziyin

11/07/2022

Table of Contents

Phenomenon of Collapse

- There are 3 types of collapses
  - Neural collapse in supervised learning
  - Posterior collapse in Bayesian deep learning
  - Dimensional collapse in self-supervised learning
- While independently discovered, all collapses happen when the learned data representation becomes low-rank
  - We thus take this as the definition of “collapse” in this presentation
    - “complete collapse”
    - “partial collapse”

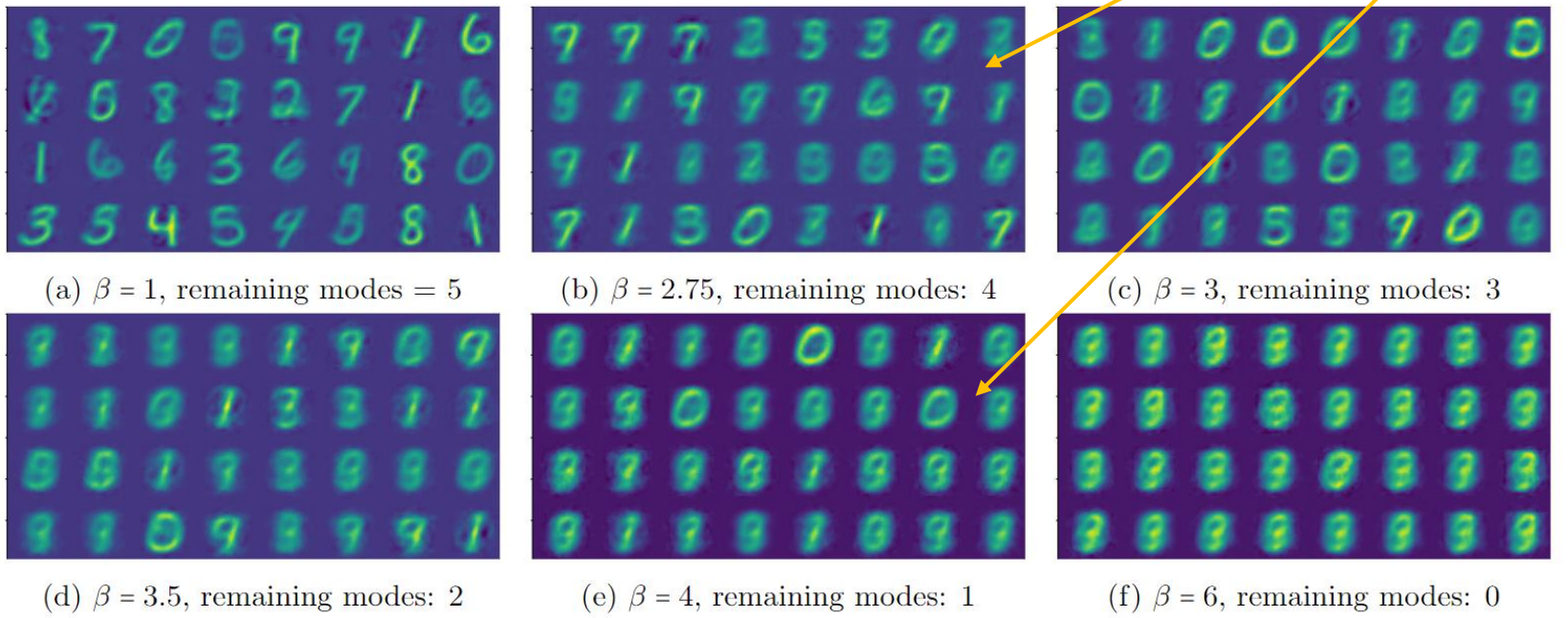
Posterior Collapse

- In Bayesian deep learning (especially in the setting of variational autoencoders (VAE)), posterior collapse happens when the learned posterior distribution becomes the same as the prior
  - Equivalent to that the mean of the latent variables is low-rank
  - Chronologically speaking, this is the earliest discovered example of collapse (Alemi et al., 2018; Lucas et al., 2019)
  - Posterior collapse is mainly regarded as a bad thing to avoid
- Objective: Minimize the loss function:

$$L = L_{likelihood} + \beta L_{prior}$$

  - Collapse happens when β is large

Posterior Collapse



(a) $\beta = 1$, remaining modes $= 5$ (b) $\beta = 2.75$, remaining modes: 4 (c) $\beta = 3$, remaining modes: 3

(d) $\beta = 3.5$, remaining modes: 2 (e) $\beta = 4$, remaining modes: 1 (f) $\beta = 6$, remaining modes: 0

Figure 3: MNIST generation under different β. We see that the generated images lose diversity and variation as β increases. The number of mode left is estimated by the theoretical prediction of thresholds of each singular values.

Posterior Collapse

- An open question:

  Why does posterior collapse only happen for VAE, but not for other types of Bayesian learning?

Neural Collapse

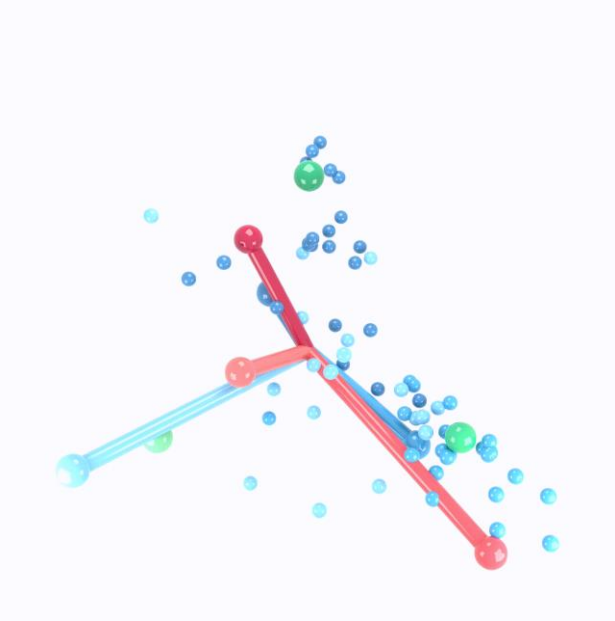
- In supervised learning, a neural collapse happens when the data representations are the same as the class mean
  - Namely, it happens when the inner class variation vanishes (Papyan 2020)
  - Neural collapse is commonly regarded as a good thing because it suggests that the model is only learning the task relevant features
- Objective: Minimize the loss function:

$$L = L_{classification} + \gamma[L_2\, reg.]$$

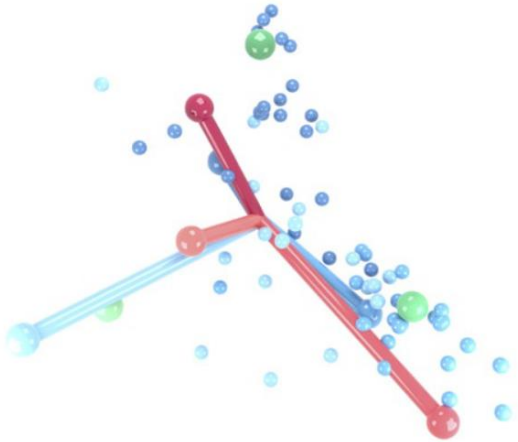
Neural Collapse

- Objective: Minimize the loss function:

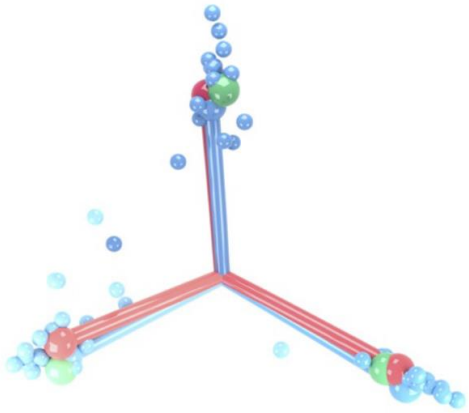
$$L = L_{classification} + \gamma[L_2\, reg.]$$



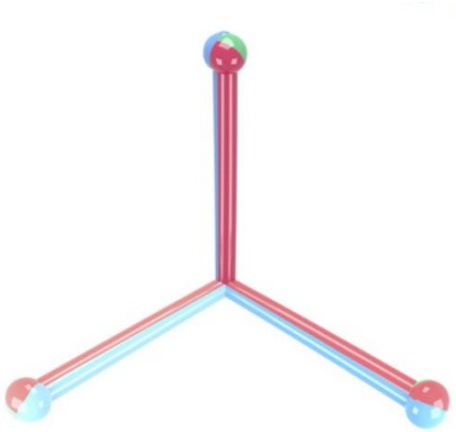
Init.



(roughly) zero training error.



convergence



Neural Collapse

- Why does the collapse not happen for a linear regressor?
- Ridge linear regressor:

$$w^* = [E[xx^T] + \gamma I]^{-1} E[xy]$$

Dimensional Collapse

- In self-supervised learning, dimensional collapse refers to the case when the learned representation becomes low-rank (Jing et al, 2021)
  - Sometimes, it is regarded as a good thing (Cosentino et al., 2022), sometimes it is regarded as a bad thing (Jing et al, 2021)

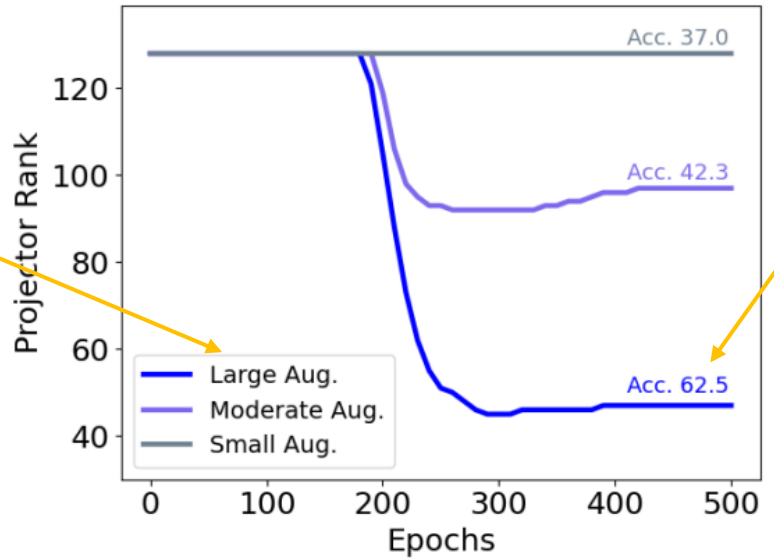
- Objective: Minimize the loss function:

$$L = L_{attraction} + L_{repulsion}(\alpha)$$

  - $L_{attraction}$ encourages representations of similar data to be close
  - $L_{repulsion}$ encourages representations of dissimilar data to be distant
  - α controls the relative tradeoff between $L_{attraction}$ and $L_{repulsion}$
    - One example is the strength of **data augmentation**

Dimensional Collapse

A stronger augmentation leads to lower rank



A lower rank leads to better performance

Figure 1: **Evolution throughout contrastive SSL training of the rank of a linear projector of dimension 512×128 for different augmentation strengths, and the associated accuracy obtained on Cifar100 by using the representation extracted in the encoder space.** Large, moderate, and small augmentations refer to the strength of the data augmentation applied to the input samples (see Table 2 for each configuration). The smaller the strength of the data augmentation policy, the less the projector suffers from dimensional collapse. However, when the projector is affected by a substantial dimensional collapse, the encoder representation becomes suitable for the downstream task. In this work, we demystify this intriguing relationship between augmentation strengths, encoder embedding, and projector geometry.

Collapses

- In short, *collapses happen everywhere in deep learning*
  - Why?
  - Is there a universal cause?

- These examples often share two common features
  1. A data-dependent term and an (effective) regularization term exist
     - A Natural Hypothesis: competition between data learning and regularization effect leads to collapses
  2. Models are often trained for very long and sometimes to convergence
     - A Natural Hypothesis: the collapses are related to the properties of the stationary points of the training objective

Table of Contents

A Landscape Perspective of Collapse

- Is the competition between feature learning and regularization sufficient to cause collapse?
  - No.

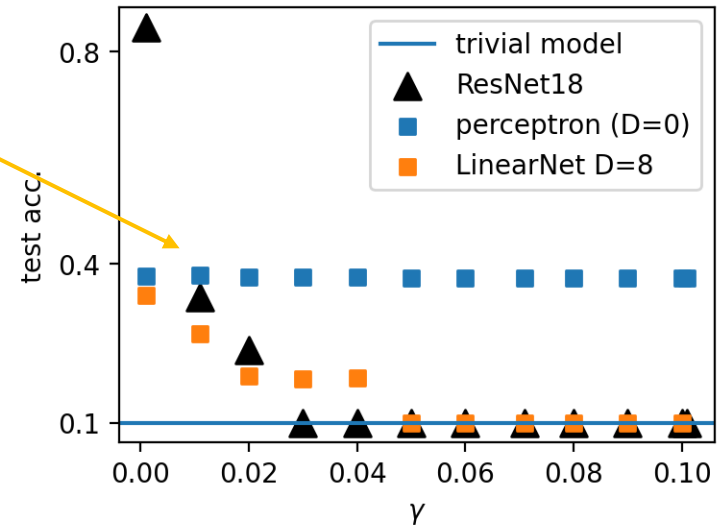
A Landscape Perspective of Collapse

- Minimal example:
  - Ridge linear regression: $L(W) = \mathrm{E}_x \left|\left|Wx - y\right|\right|^2 + \gamma \left|\left|w\right|\right|^2$
  - Let $A_0 \coloneqq \mathrm{E}_x[xx^T]$ denote the feature covariance (or, just, "feature" for short)
  - The stationary point is unique:

$$W = [A_0 + \gamma I]^{-1} E_x[xy]$$

- The model is full-rank as along as A_0 is full-rank -- there is no collapse
- Lesson: we need either an advanced loss function or a deeper model

Linear regression is insensitive to regularization, whereas deep models are hypersensitive



A Landscape Perspective of Collapse

- Hypothesis: *Depth plays a key role in collapse*
- Minimal example (one-h-layer linear model):

$$L_{d,d_1}(U,W) = \mathbb{E}_x \mathbb{E}_\epsilon \left(\sum_j^{d_1} U_j \epsilon_j \sum_i^{d} W_{ji} x_i - y \right)^2 + \gamma_w ||W||^2 + \gamma_u ||U||^2,$$

  - $y = y(x)$ is one-dimensional
  - ϵ are independent random variables with unit mean and σ^2 variance
    - For example, due to the use of dropout
  - Two layers have different strengths of regularization
  - We want to find the global minimum W^* and U^*

Solution

- Let b denote the norm of the model ($b = \sqrt{\left|\left|W\right|\right|^2 + \left|\left|U\right|\right|^2}$)
- One can show that, defining $t := \left|\left|E_x[xy]\right|\right| - \sqrt{\gamma_u \gamma_w}$, at global minima

$$b \propto \sqrt{t} \quad if \quad t > 0$$
$$b = 0 \quad if \quad t \leq 0$$

No collapse

Complete collapse

- Some "physics" messages:
  1. b is an order parameter
  2. We have a second-order phase transition (if we treat the training loss as a free energy)
  3. $\left|\left|E_x[xy]\right|\right|^2 = \gamma_u \gamma_w$ is the critical point

We can prove rigorously…

Theorem 1. *The global minimum U_* and W_* of Eq. (2) is $U_* = 0$ and $W_* = 0$ if and only if*

$$||\mathbb{E}[xy]||^2 \leq \gamma_u \gamma_w. \tag{4}$$

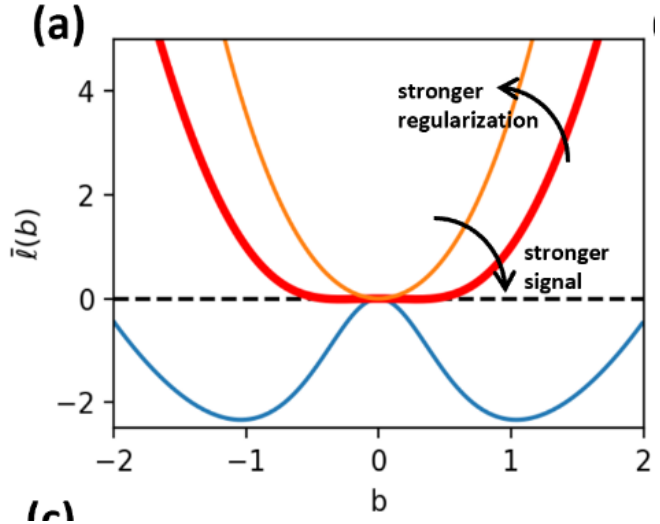
When $||\mathbb{E}[xy]||^2 > \gamma_u \gamma_w$, the global minima are

$$\begin{cases} U_* = b\mathbf{r}; \\ W_* = \mathbf{r}\mathbb{E}[xy]^T b \left[b^2\left(\sigma^2 + d_1\right) A_0 + \gamma_w I\right]^{-1}, \end{cases} \tag{5}$$

where $\mathbf{r} = (\pm 1, ..., \pm 1)$ is an arbitrary vertex of a d_1-dimensional hypercube, and b satisfies:

$$\left\| \left[b^2\left(\sigma^2 + d_1\right) A_0 + \gamma_w I\right]^{-1} \mathbb{E}[xy] \right\|^2 = \frac{\gamma_u}{\gamma_w}. \tag{6}$$

Effective loss landscape:



A Landscape Perspective of Collapse

- Lesson: *complete collapse happens in a two-layer linear model at*

$$\left|\left|E_x[xy]\right|\right| = \gamma_u\gamma_w$$

- This critical point is rather universal because
  - Independent of width or data dimension
  - Independent of the data covariance
  - Independent of the noise
- Interpretation: a collapse happens due the competition between signal strength and regularization

Posterior Collapse

- For Bayesian deep learning, the minimal model is a linear latent variable model
- Data generation process: $x \to z \to y$
  - When $y = x$, we have an autoencoder
- Loss function:

$$\mathbb{E}_x[-\mathbb{E}_{q(z|x)} \log(p(y|z)) + \beta D_{KL}(q(z|x) \| p(z; \eta_{\text{enc}}^2))]$$

- In case of linear encoder and decoder, and Gaussian assumption, the loss function is (U is the decoder, W is the encoder)

$$\frac{1}{2\eta_{\text{dec}}^2} \mathbb{E}_{x,\epsilon} \left[\| U(W^\top x + \epsilon) - y \|^2 + \beta \frac{\eta_{\text{dec}}^2}{\eta_{\text{enc}}^2} \| W^\top x \|^2 \right] + \sum_{i=1}^{d_1} \frac{\beta}{2} \left(\frac{\sigma_i^2}{\eta_{\text{enc}}^2} - 1 - \log \frac{\sigma_i^2}{\eta_{\text{enc}}^2} \right)$$

Posterior Collapse

- The singular values (λ_i, θ_i) of U and W are the order parameters, and the phase transition is again second-order

$$\lambda_i \propto \sqrt{\zeta_i^2 - \beta\eta_{dec}^2}$$
$$\lambda_i = 0$$

- ζ_i^2 are the eigenvalues of $E_x[xy]\mathrm{E}_x[xy]^T$
- η_{dec}^2 is the prior variance of the decoder

Posterior Collapse

- As before, we can find the global minimum rigorously

Theorem 2. *The global minimum of* $L_{\mathrm{VAE}}(U, W, \Sigma)$ *is given by*

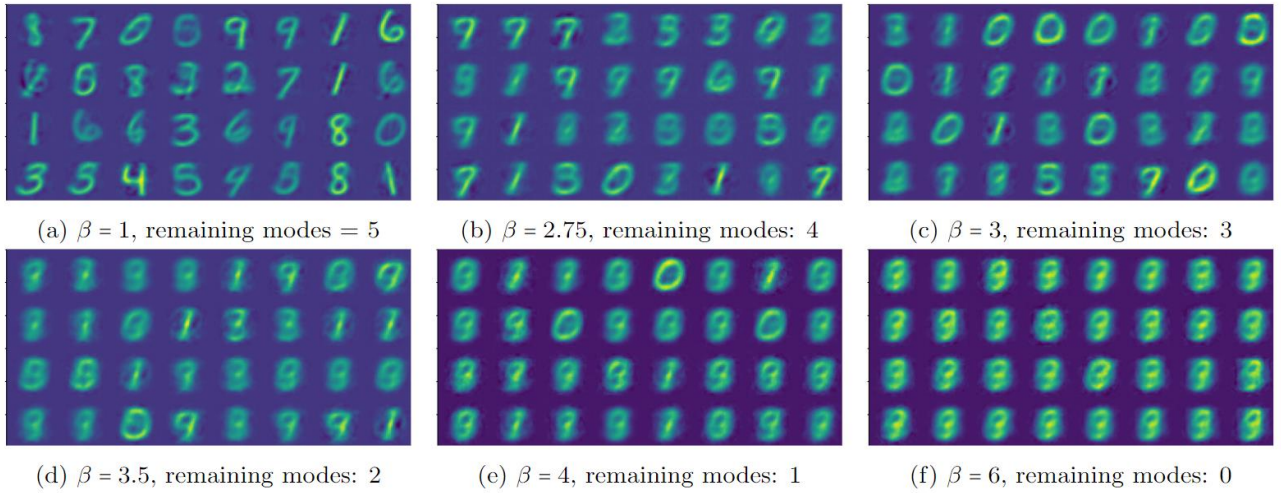
$$U^* = F\Lambda P, \quad W^* = A^{-\frac{1}{2}} G\Theta P, \tag{16}$$

where F *and* G *are derived by the SVD of* Z, P *is an arbitrary orthogonal matrix in* $\mathbb{R}^{d_1 \times d_1}$, *and* $\Lambda = \mathrm{diag}(\lambda_1, ..., \lambda_{d_1})$ *and* $\Theta = \mathrm{diag}(\theta_1, ..., \theta_{d_1})$ *are diagonal matrices such that*

$$\lambda_i = \frac{1}{\eta_{\mathrm{enc}}} \sqrt{\max\left(0, \zeta_i^2 - \beta\eta_{\mathrm{dec}}^2\right)} \tag{17}$$

$$\theta_i = \frac{\eta_{\mathrm{enc}}}{\zeta_i} \sqrt{\max\left(0, \zeta_i^2 - \beta\eta_{\mathrm{dec}}^2\right)}. \tag{18}$$

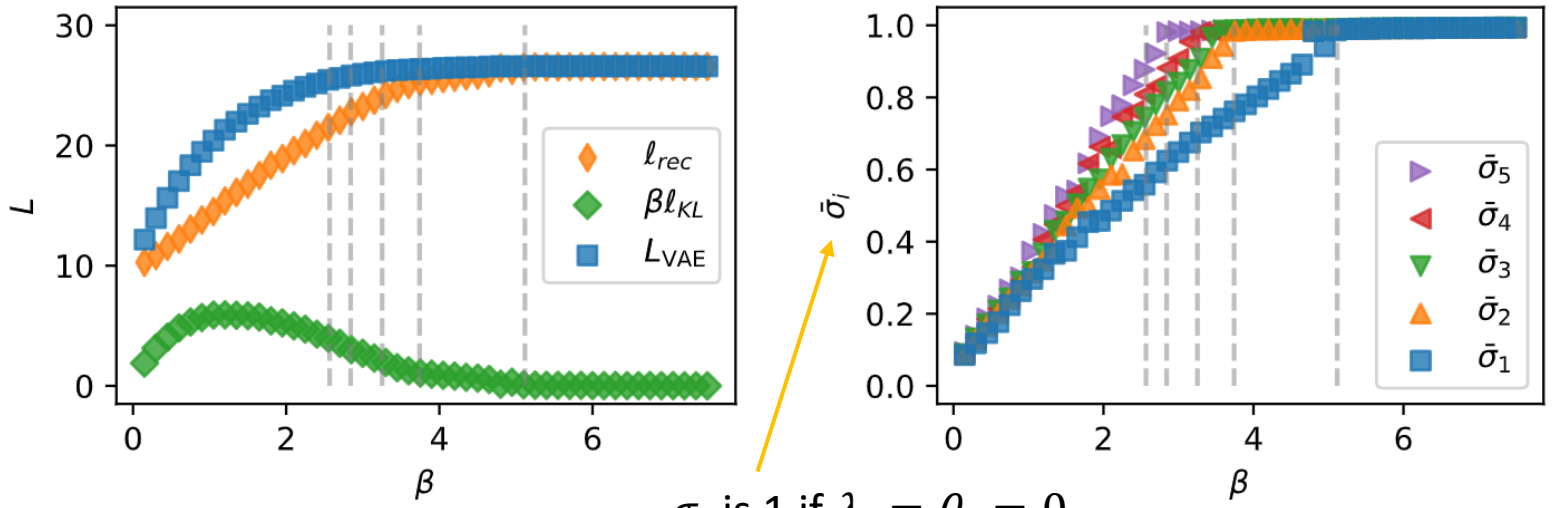
Posterior Collapse



(a) $\beta = 1$, remaining modes $= 5$ (b) $\beta = 2.75$, remaining modes: 4 (c) $\beta = 3$, remaining modes: 3

(d) $\beta = 3.5$, remaining modes: 2 (e) $\beta = 4$, remaining modes: 1 (f) $\beta = 6$, remaining modes: 0

- The singular values (λ_i, θ_i) of U and W are the order parameters, and the phase transition is again second-order

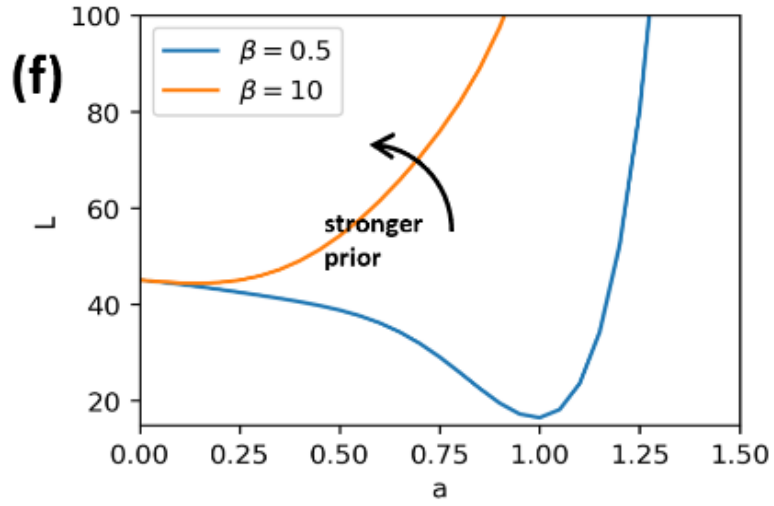


σ_i is 1 if $\lambda_i = \theta_i = 0$

- There are multiple collapses (equal to the number of distinct ζ_i)

Posterior Collapse

- We can also plot the effective landscape
  - Train a model on MNIST in the non-collapsed phase and rescale all the weights by a



Dimensional Collapse

- Dimensional collapses are a little different: *it does not require depth*
- The most standard loss function InfoNCE can be written as

$$L_\epsilon = \mathbb{E}_{\hat{x}}\left\{\frac{1}{2}|f(x) - f(x')|^2 + \log \mathbb{E}_{\hat{\chi}}\left[\exp\left(-\frac{1}{2}|f(x) - f(\chi)|^2\right)\right]\right\},$$

1. Sample a data $\hat{x}$

2. Add data augmentation ϵ to generate x and x'

3. Sample a different data $\hat{\chi}$, and apply data augmentation

- Encourages x and x' to be similar
- Encourages x and χ to be dissimilar
- The loss function is invariant to a simultaneous rotation of the learned representation: $f(x) \to Rf(x)$

Dimensional Collapse

- Linear model: $f(x) = Wx$
- Expand the log-exponential term to fourth order
- Assume that the data is Gaussian

Landau theory!

$$L = -\mathrm{Tr}[WBW^T] + \mathrm{Tr}[W\Sigma W^T W\Sigma W^T].$$

Noise strength

where B is a linear function of A_0 and C, and $\Sigma = A_0 + C$

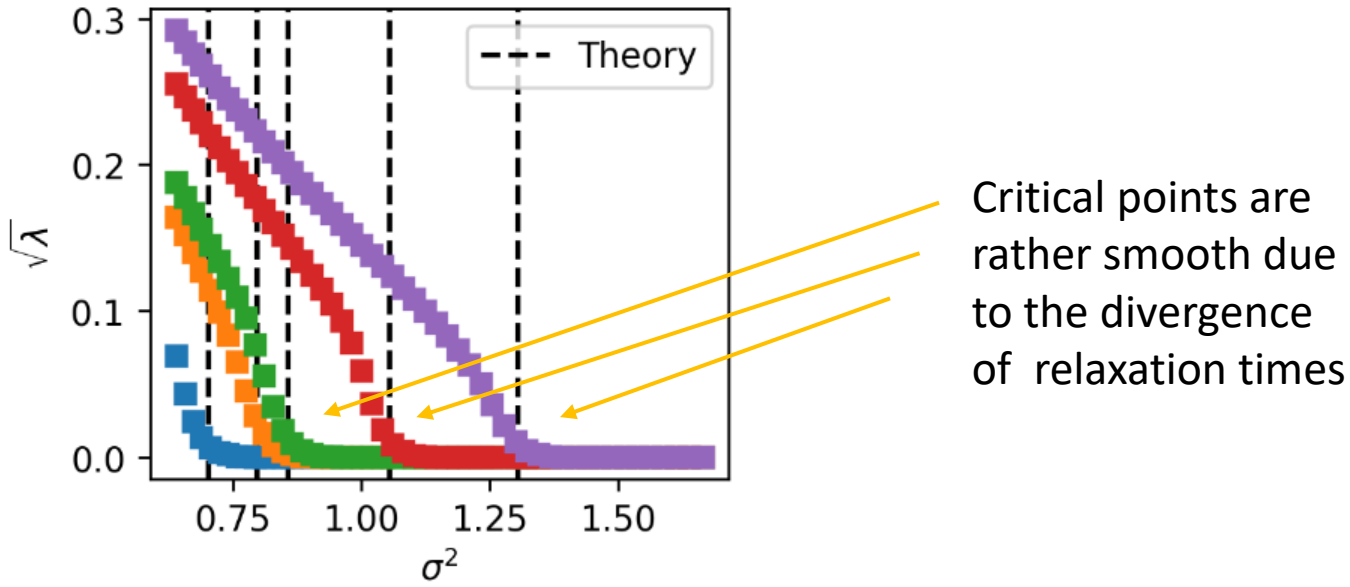
Feature variance

- Note: a crucial feature of this loss is that the odd-order terms vanish. Almost all SSL losses can reduce to this form because of the rotational symmetry

Dimensional Collapse

- The square root of the eigenvalues of WW^T are the order parameters
- When $C = \sigma^2 I$

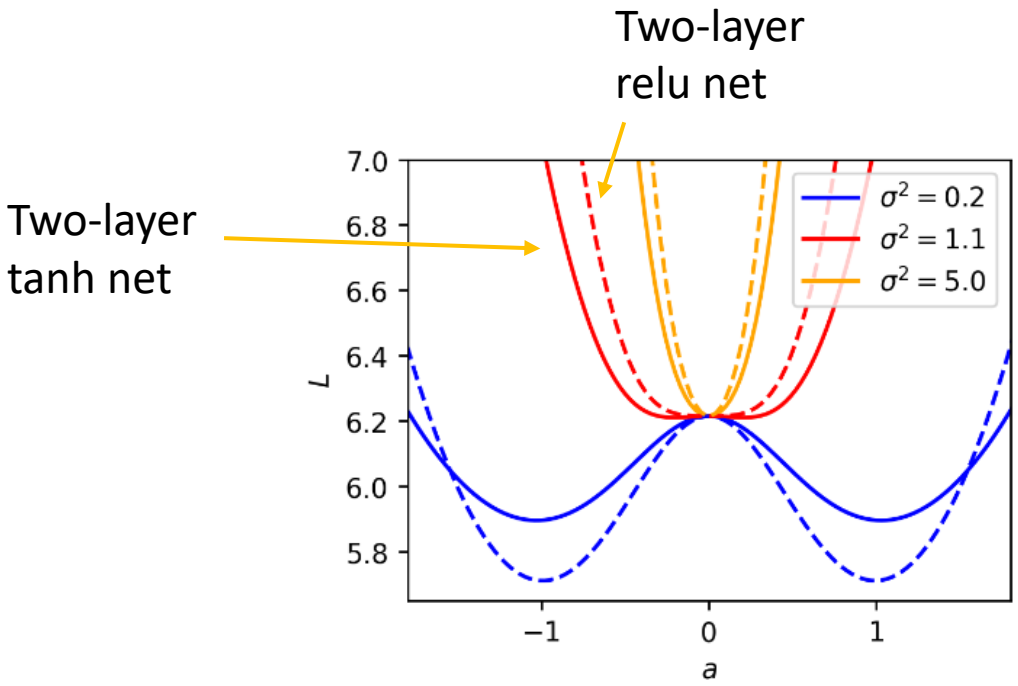
$$\sqrt{\lambda_i} \propto \sqrt{a_i - \sigma^2}$$



Critical points are rather smooth due to the divergence of relaxation times

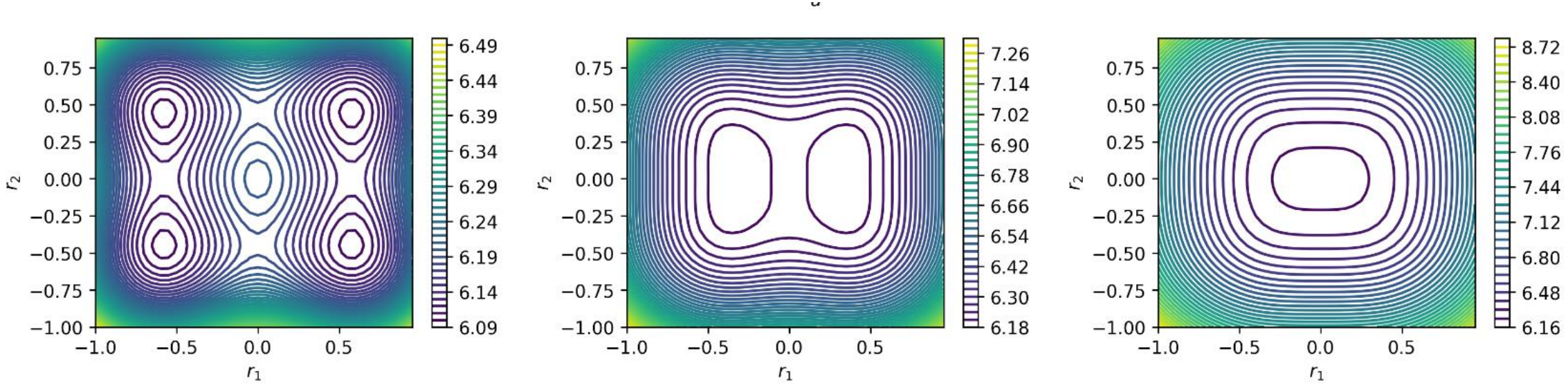
Dimensional Collapse

- The effective landscape is similar for nonlinear models:



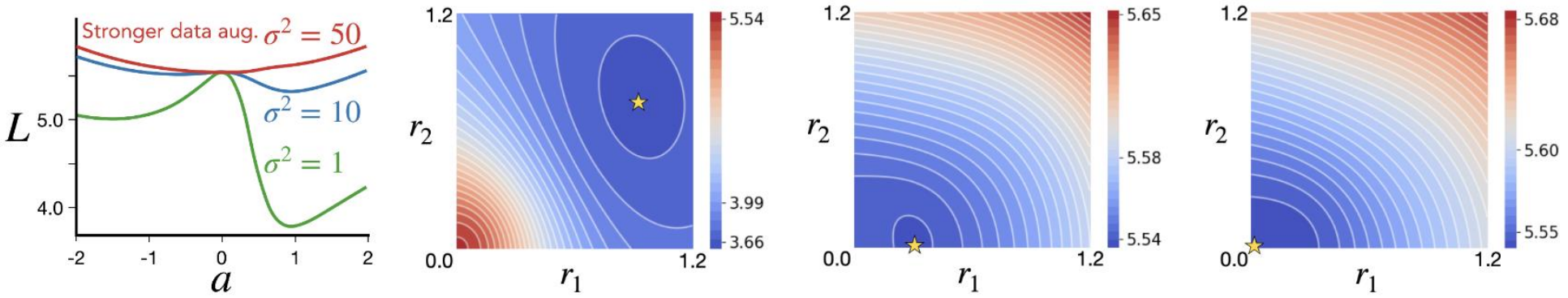
Dimensional Collapse

- 2-d projection of the ReLU net landscape
- $f(x) \in R^2$
- We rescale the two rows of the last-layer weight matrix by r_1 and r_2 respectively



Dimensional Collapse

- ResNet18 on CIFAR-10:



(a) 1D landscape (b) No collapse (c) Dimensional collapse (d) Complete collapse

Figure 2: **Landscape of Resnet18 on CIFAR10 with SimCLR qualitatively agrees with our linear theory.** (a) Training objective L as a function of a rescaling of the last layer $W \to aW$. (b-d) L as a function of a $2d$ rescaling of the last layer where the data augmentation strength is (b) small, (c) intermediate, and (d) strong. Red indicates areas of high loss, blue indicates areas of low loss, and stars locate local minima. The use of data augmentation changes the stability of the origin, a qualitative change that leads to different types of collapses in qualitative agreement with our linear theory (cf. Figure 1). Additionally, we also notice the same qualitative changes in the landscape in simpler nonlinear models (see Appendix A).

“Benignity” Theorems

- So far, we only considered the global minimum of the landscape
- Can local minimum cause collapse?
  - In many cases, no.

“Benignity” Theorems

- All local minima achieve the maximum possible rank
  - Corollary: all local minima have the same rank
  - Applies to posterior collapse, dimensional collapse, and supervised learning with 1 hidden layer
  - So the collapse phenomenon is something rather independent of initialization or dynamics
- The saddle points all have lower rank than the local minima
  - Unless the model is converging to saddles, analysis of the global minimum is sufficient to understand collapse
- Lesson: the landscape of deep learning can be *rather* benign

Do we only have second-order phase transitions?

- No.

Supervised Learning

- Recall our two-layer linear model for supervised learning:

$$L_{d,d_1}(U,W) = \mathbb{E}_x \mathbb{E}_\epsilon \left(\sum_j^{d_1} U_j \epsilon_j \sum_i^d W_{ji} x_i - y \right)^2 + \gamma_w ||W||^2 + \gamma_u ||U||^2,$$

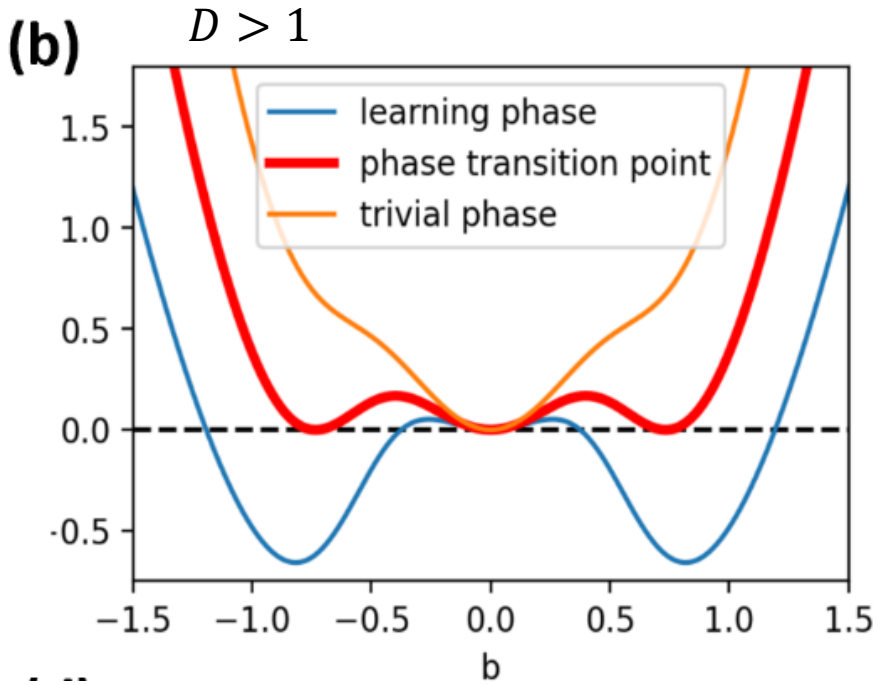
- We can generalize it to multiple layers:

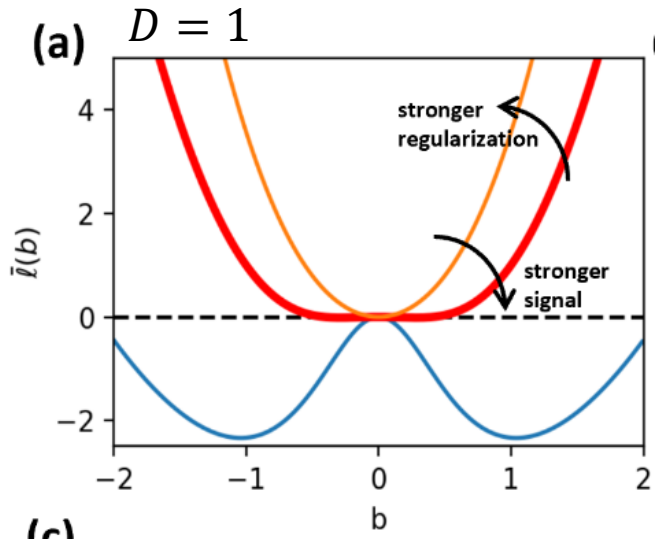
$$\mathbb{E}_x \mathbb{E}_{\epsilon^{(1)},\epsilon^{(2)},\ldots,\epsilon^{(D)}} \left(\sum_{i,i_1,i_2,\ldots,i_D}^{d,d_1,d_2,\ldots d_D} U_{i_D} \epsilon_{i_D}^{(D)} \ldots \epsilon_{i_2}^{(2)} W_{i_2 i_1}^{(2)} \epsilon_{i_1}^{(1)} W_{i_1 i}^{(1)} x_i - y \right)^2 + \gamma_u ||U||^2 + \sum_{i=1}^{D} \gamma_i ||W^{(i)}||^2,$$

  - D: the number of hidden layers

Supervised Learning

- Effective landscape: $$\bar{\ell}(b,\gamma) := -\sum_i \frac{d_0^{2D} b^{2D} \mathbb{E}[x'y]_i^2}{d_0^D(\sigma^2+d_0)^D a_i b^{2D}+\gamma} + \mathbb{E}_x[y^2] + \gamma D d_0^2 b^2,$$





- $b = 0$ is always a local minimum for $D > 1$

Supervised Learning

- There is a first-order phase transition whenever $D > 1$.
- $b(\gamma)$ features a jump at a critical value of γ

Theorem 2. *Any global minimum of Eq. (9) is of the form*

$$\begin{cases} U = b_u \mathbf{r}_D; \\ W^{(i)} = b_i \mathbf{r}_i \mathbf{r}_{i-1}^T; \\ W^{(1)} = \mathbf{r}_1 \mathbb{E}[xy]^T (b_u \prod_{i=2}^{D} b_i) \mu \left[(b_u \prod_{i=2}^{D} b_i)^2 s^2 \left(\sigma^2 + d_1\right) A_0 + \gamma_w I \right]^{-1}, \end{cases} \tag{10}$$

where $\mu = \prod_{i=2}^{D} d_i$, $s^2 = \prod_{i=2}^{D} d_i(\sigma^2 + d_i)$, $b_u \geq 0$ *and* $b_i \geq 0$, *and* $\mathbf{r}_i = (\pm 1, \ldots, \pm 1)$ *is an arbitrary vertex of a* d_i*-dimensional hypercube for all* i. *Furthermore, let* $b_1 := \sqrt{\|W_{i:}\|^2/d}$ *and* $b_{D+1} := b_u$, b_i *satisfies*

$$\gamma_{k+1} d_{k+1} b_{k+1}^2 = \gamma_k d_{k-1} b_k^2. \tag{11}$$

A Tentative Definition of Phase Transition

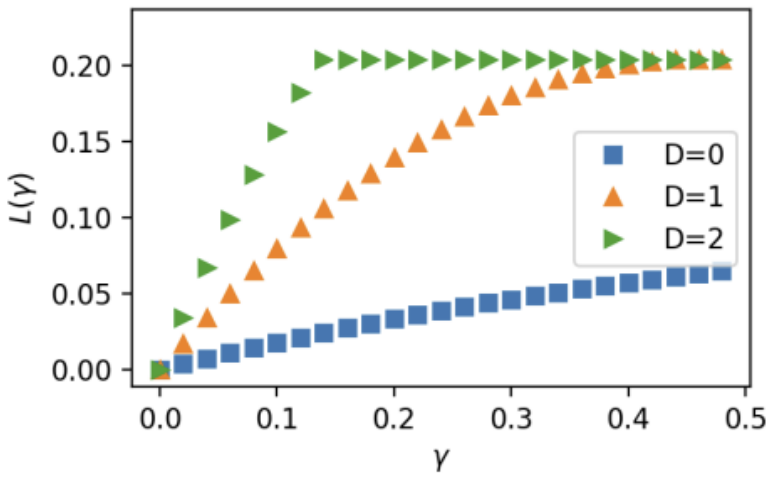
Definition (Ehrenfest-type phase transition):

$$L^*(\gamma) \coloneqq \min_W L(W, \gamma)$$

We say that n-th order phase transition happens if $\left(\frac{d}{d\gamma}\right)^n L(\gamma)$ is discontinuous

Phase Transition

- One can prove the following results:
  1. There is no zeroth-order phase transition (PT) for finite D
  2. $D = 0$ has no PT
  3. $D = 1$ has a second-order PT
  4. $D \geq 2$ has a first-order PT



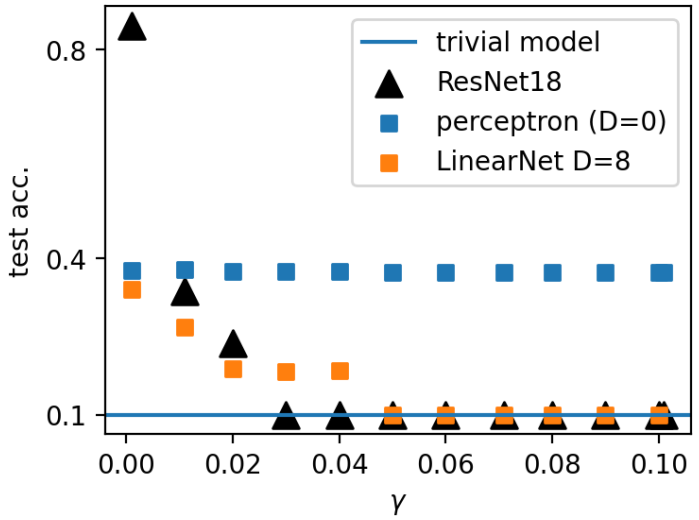
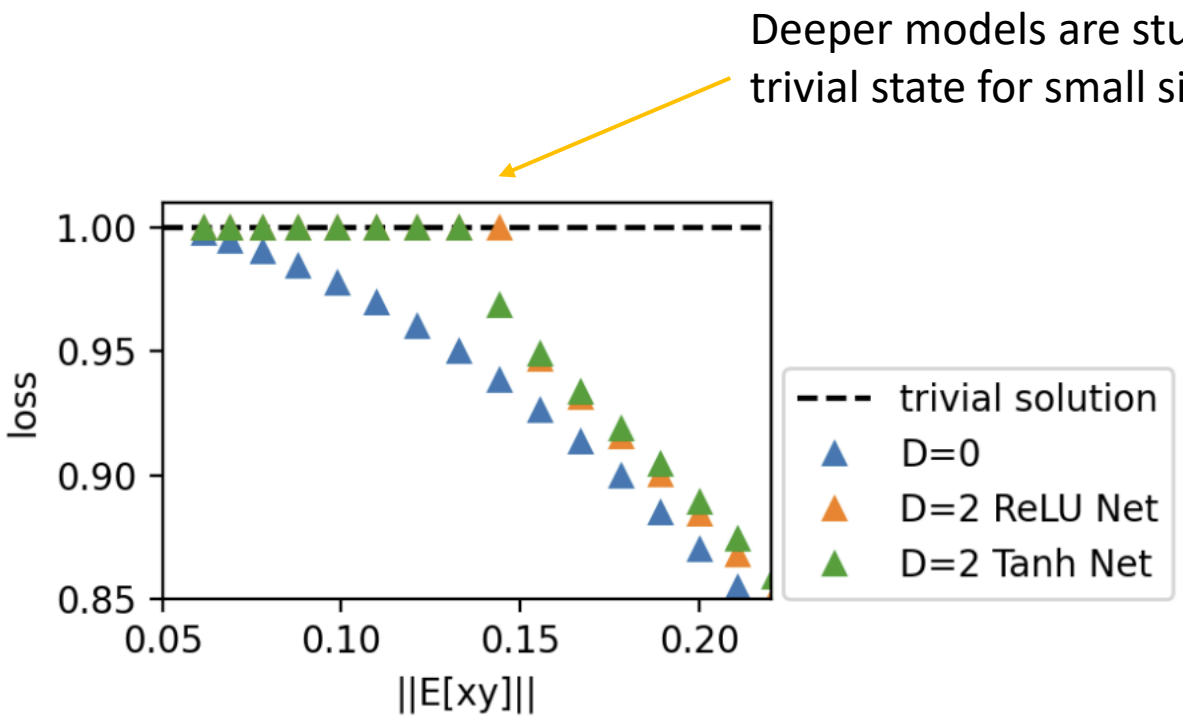


Table of Contents

Implications

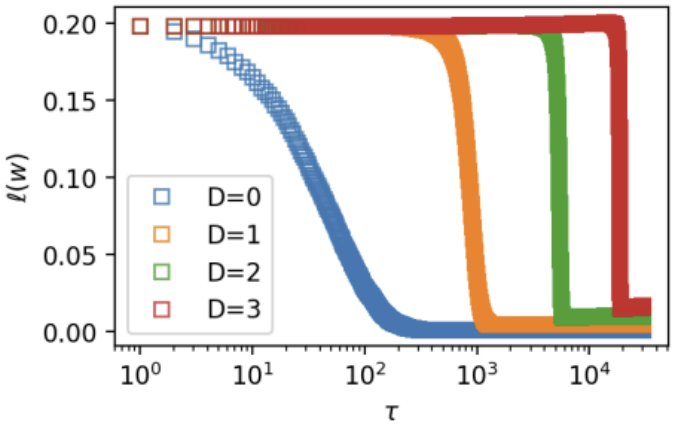
- Need to escape local minimum for deeper models



Implications

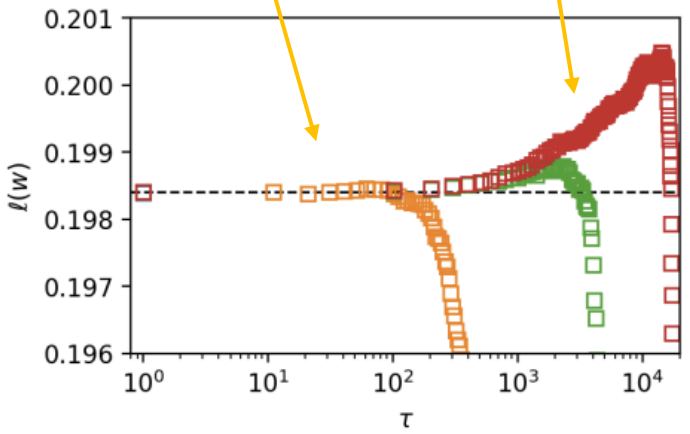
- Critical dynamics takes a rather universal characteristic
- Training proceeds with gradient descent plus additive Gaussian noise

Deeper models feature a wide plateau at initialization

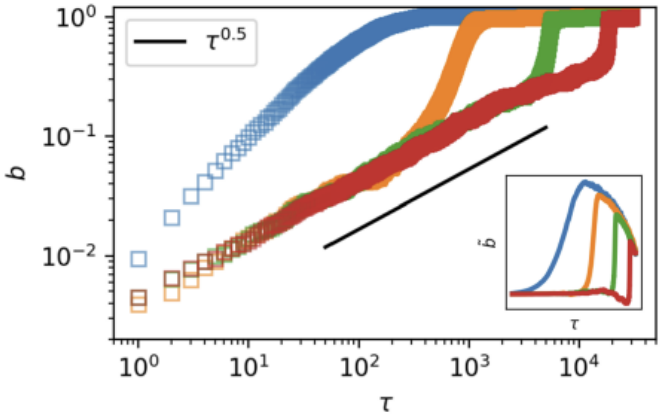


$D = 1$: no barrier to overcome

$D > 0$: barriers to overcome



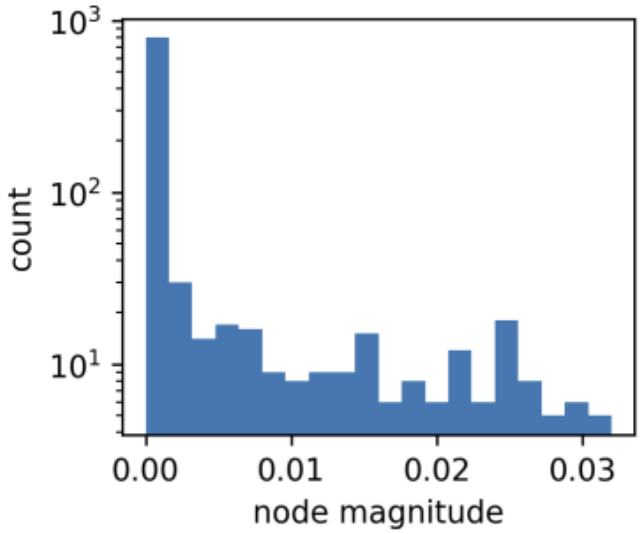
Learning of deeper models are divided into regimes

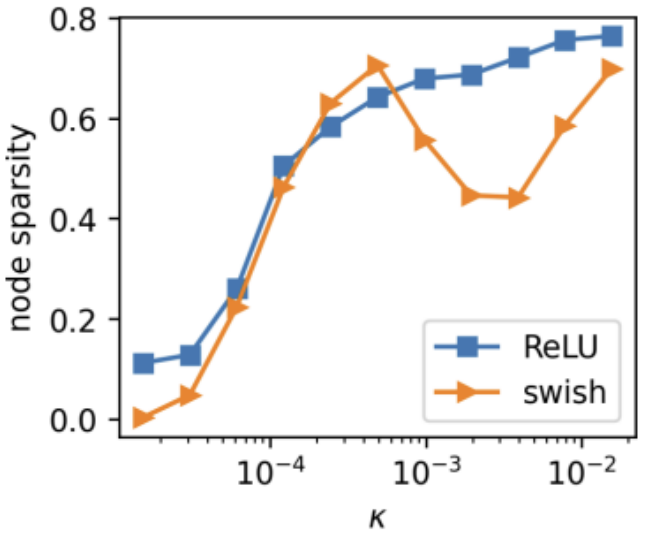


Implications

- Sparsity is common in deep regularized models

- Example of a two-layer model trained on MNIST
  - κ: weight decay





Last Comment

- Is landscape the only cause of collapse?
  - No.

Last Comment

- Besides the landscape causes of collapse, SGD dynamics can also lead to collapse (implicit regularization).
  - Complete collapse can happen when the learning rate is too large and when the batch size is too small

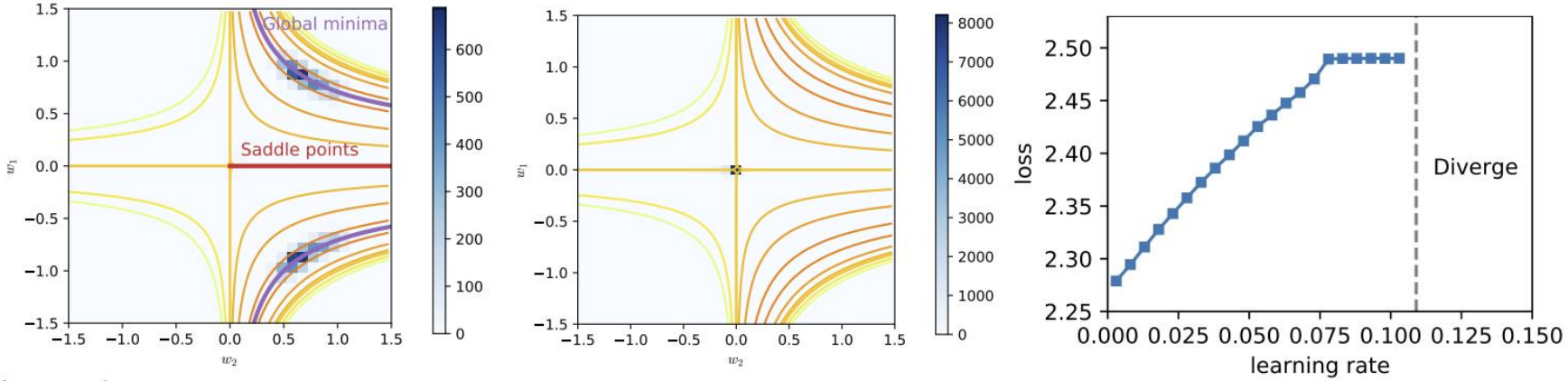
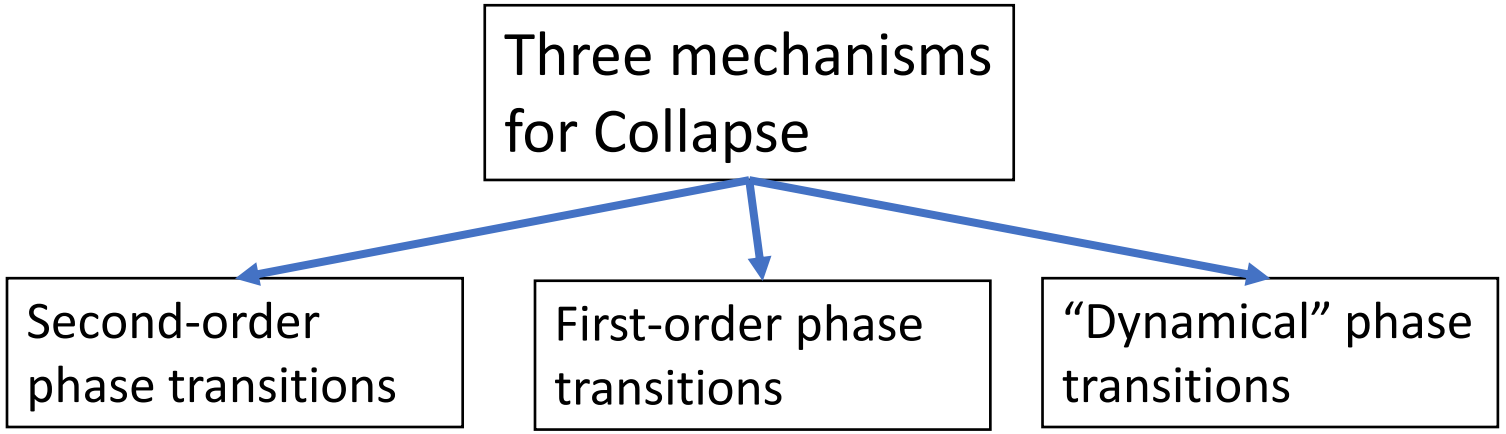


Figure 3: Convergence of a two-layer one-neuron neural network to a saddle point. The blue region shows the empirical density of converged parameter distribution. **Left**: $\lambda = 0.001$ at step 10000 converges to global minima. **Mid**: $\lambda = 0.1$ at step 10000 converges to a saddle point. **Right**: Average loss in equilibrium as a function of learning rate. The loss function diverges for learning rates larger than 0.108.

A Unifying Picture



Stability of the origin matters

- *Posterior collapse*
- *Dimensional collapse*
- *1-hidden layer models*

Difficult to understand

- *Occur in deeper models*
- *Factors are global*
- *Need to overcome loss barriers*

Induced by the implicit bias of SGD

- *Large learning rate*
- *Small batch size*
- *Should occur in any model (?)*

Messages

- Collapse is a ubiquitous phenomenon in deep learning and there might exist universal explanation for it
- Landscape analysis around the origin can explain the landscape causes of the collapse phenomenon
  - The origin is a very special point in deep learning!
- Phase transition behaviors are ubiquitous in deep learning in the form of collapses
- (Alternatively) Collapses can be understood in the form of phase transitions
  - Norm of the model / singular values of the weight matrices are good candidates of order parameters
- There are also non-landscape causes of collapse
  - Implicit bias!

An Important Question

- *Can collapse explain the success of deep learning*?
  - Collapse encourages low-rankness and sparsity, which are good candidates for explaining generalization
  - Can we use collapse to design better learning algorithms?
  - Do biological brains “collapse”?

Partially based on…

- “Exact Solutions of a Deep Linear Network.” NeurIPS 2022
- “Posterior Collapse of a Linear Latent Variable Model.” NeurIPS 2022
- “Exact Phase Transitions in Deep Learning.” arxiv 2205.12510
- “What shapes the loss landscape of self-supervised learning?” arxiv 2210.00638
- “SGD with a Constant Large Learning Rate Can Converge to Local Maxima.” ICLR 2022

Collaborators: *Masahito Ueda, Hidenori Tanaka, Ekdeep Singh Lubana, Botao Li, Zihao wang, James B. Simon, Xiangming Meng*

End of Presentation